Risk and Portfolio Management
Spring 2011

Principal Components Analysis and

Factors explaining stock returns

Correlation matrices in finance

$$\Gamma_{ij} = Corr\left(R_i, R_j\right) \qquad R_i = \text{return of stock } \# i,\ \ i = 1,\ldots,N$$

Estimation of correlation matrix from data requires selecting a sample size, or estimation window, *T.*

If the universe of assets is large, then *T<<N (e.g. T=252, N=1500)*

The correlation matrix is not ``full rank'' in general since we expect that the stocks are ``driven'' by a few components

$$R_i = \alpha_i + \sum_{k=1}^{m} \beta_{ik} F_k + \varepsilon_i, \quad m << N, \ \left(e.g., \ m = 15\right)$$

- Degeneracy issue (rank<dimension)
- Noise issue (determine the ``right'' number of factors, avoid numerical error)

Principal components analysis of stock returns

-- Define a universe, or collection of stocks corresponding to the market of interest (e.g. US Equities, Nasdaq-100, Brazilian equities components of S&P 500)

-- Collect as much data as possible

-- On any given date, perform PCA on the correlation matrix, going back for T periods (days). The analysis is on a T by N matrix

-- Estimate the number of significant components

-- Analyze the corresponding eigenvectors and eigenportfolios (factors)

-- Associate the factors to features of the market (e.g. sectors, market cap, etc)

``Clean'' correlation matrix from factor model

$$R_i = \alpha_i + \sum_{k=1}^{m} \beta_{ik} F_k + \varepsilon_i, \quad m << N, \ (e.g., \ m = 15)$$

$$\Gamma_{ij} = \sum_{kl=1,m} \beta_{ik} G_{kl} \beta_{jl} + D_{ij} = \underbrace{\Gamma_{ij}^{(S)}}_{\text{Systematic correlation}} + \underbrace{\Gamma_{ij}^{(I)}}_{\text{Idiosyncratic correlation}}$$

- Systematic component is N by N with rank m
- Idiosyncratic component is diagonal N by N

This approach decomposes the variance of any asset into
a sum of ``market-explained'' variance and a company-specific variance

Stocks of more than 1BB cap in January 2007

Sector	ETF	Num of Stocks	Market Cap		unit: 1M/usd
			Average	Max	Min
Internet	HHH	22	10,350	104,500	1,047
Real Estate	IYR	87	4,789	47,030	1,059
Transportation	IYT	46	4,575	49,910	1,089
Oil Exploration	OIH	42	7,059	71,660	1,010
Regional Banks	RKH	69	23,080	271,500	1,037
Retail	RTH	60	13,290	198,200	1,022
Semiconductors	SMH	55	7,303	117,300	1,033
Utilities	UTH	75	7,320	41,890	1,049
Energy	XLE	75	17,800	432,200	1,035
Financial	XLF	210	9,960	187,600	1,000
Industrial	XLI	141	10,770	391,400	1,034
Technology	XLK	158	12,750	293,500	1,008
Consumer Staples	XLP	61	17,730	204,500	1,016
Healthcare	XLV	109	14,390	192,500	1,025
Consumer discretionary	XLY	207	8,204	104,500	1,007
Total		**1417**	**11,291**	**432,200**	**1,000**

January, 2007

Principal Components Analysis of Correlation Data

Consider a time window *t=0,1,2,…,T,* (days) a universe of *N* stocks. The returns data is represented by a *T* by N matrix $\left(R_{it}\right)$

$$\sigma_i^2 = \frac{1}{T}\sum_{t=1}^{T}\left(R_{it} - \overline{R_i}\right)^2, \qquad \overline{R_i} = \frac{1}{T}\sum_{t=1}^{T} R_{it}$$

$$Y_{it} = \frac{R_{it} - \overline{R_i}}{\sigma_i}$$

Standardized returns

$$\Gamma_{ij} = \frac{1}{T}\sum_{t=1}^{T} Y_{it} Y_{jt}$$

Clearly, $Rank(\Gamma) \leq \min\left(N, T\right)$

Regularized correlation matrix

$$\Gamma_{ij} = \frac{1}{T}\sum_{t=1}^{T} Y_{it} Y_{jt} + \gamma\delta_{ij}, \quad \gamma = 10^{-9}$$

$$\Gamma_{ij}^{reg} = \frac{C_{ij}}{\sqrt{C_{ii}C_{jj}}} = \frac{C_{ij}}{1+\gamma}$$

This matrix is a correlation matrix and is positive definite. It is equivalent for all practical purposes to the original one but is numerically stable for inversion and eigenvector analysis (e.g. with Matlab).

Note: this is especially useful when *T<<N.*

Eigenvalues, Eigenvectors and Eigenportfolios

$$\lambda_1 > \lambda_2 \geq .... \geq \lambda_N > 0$$

eigenvalues

$$\mathbf{V}^{(j)} = \left(V_1^{(j)}, V_2^{(j)}, ..., V_N^{(j)}\right), \quad j = 1,2,..., N.$$

eigenvectors

$$F_{jt} = \frac{1}{\sqrt{\lambda_j}} \sum_{i=1}^{N} V_i^{(j)} Y_{it} = \sum_{i=1}^{N} \left(\frac{1}{\sqrt{\lambda_j}} \frac{V_i^{(j)}}{\sigma_i} \right) R_{it}$$

returns of "eigenportfolios"

Portfolio weights

We shall use the coefficients of the eigenvectors and the volatilities of the stocks to build ``portfolio weights''. These random variables (F_j) span same linear space as the original returns.

Expressing Stock Returns in terms of returns of Eigenportfolios (a bit of linear algebra)

$$Correl(R_i, R_j) = \Gamma_{ij} = \sum_{k=1}^{N} \lambda_k V_i^{(k)} V_j^{(k)}$$

Define: $\qquad F_k \equiv \frac{1}{\sqrt{\lambda_k}} \sum_{i=1}^{N} \frac{V_i^{(k)}}{\sigma_i} R_i$

$$Variance(F_k) = 1, \quad Cov(F_k, F_l) = \delta_{kl}$$

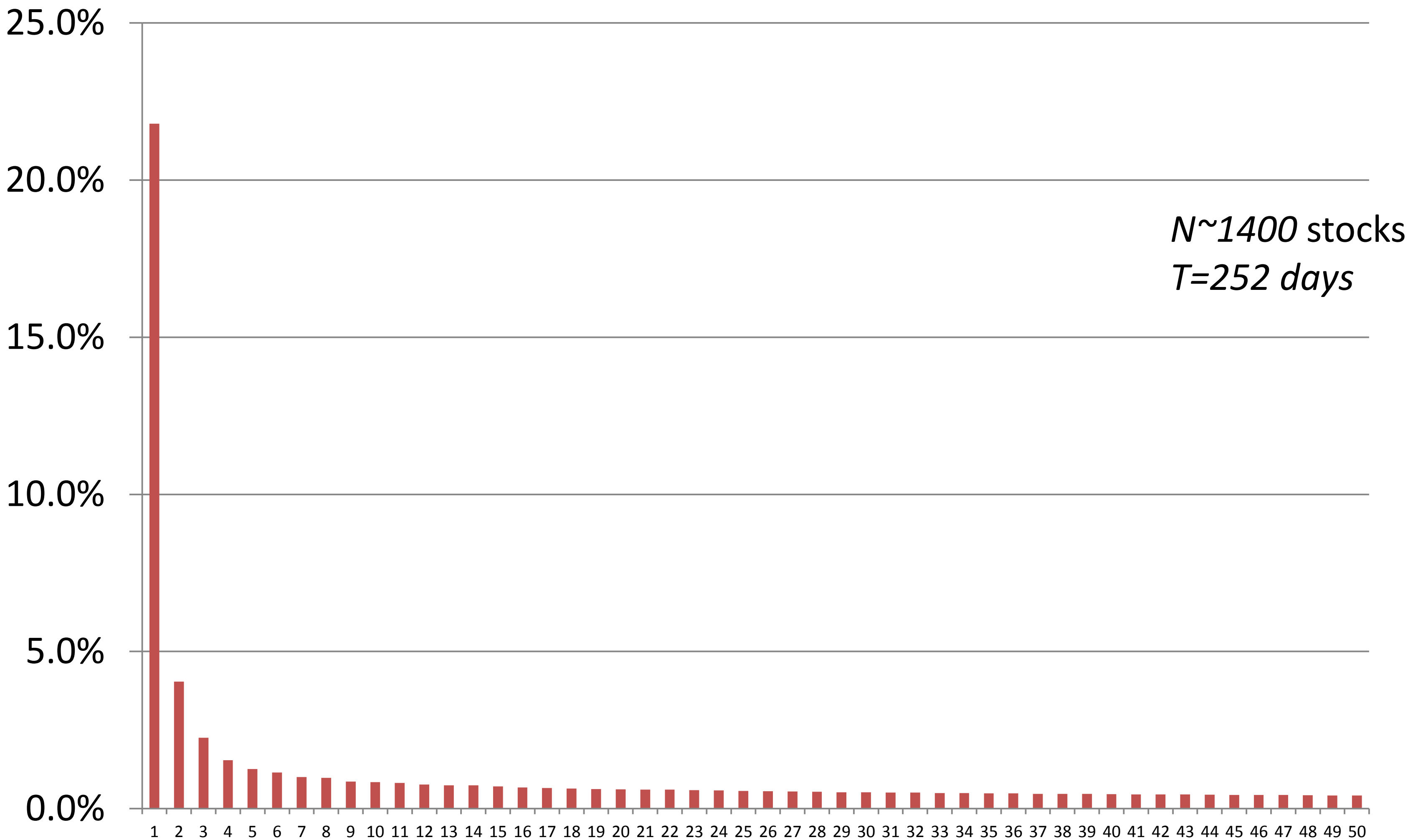
Set: $\quad \beta_{\mathrm{ik}} = Cov(R_i, F_k) = \frac{1}{\sqrt{\lambda_k}} \sum_{j=1}^{N} V_j^{(k)} \, \sigma_i \Gamma_{ij}$

$$= \frac{1}{\sqrt{\lambda_k}} \sigma_i \lambda_k V_i^{(k)} = \sigma_i \sqrt{\lambda_k} V_i^{(k)}$$

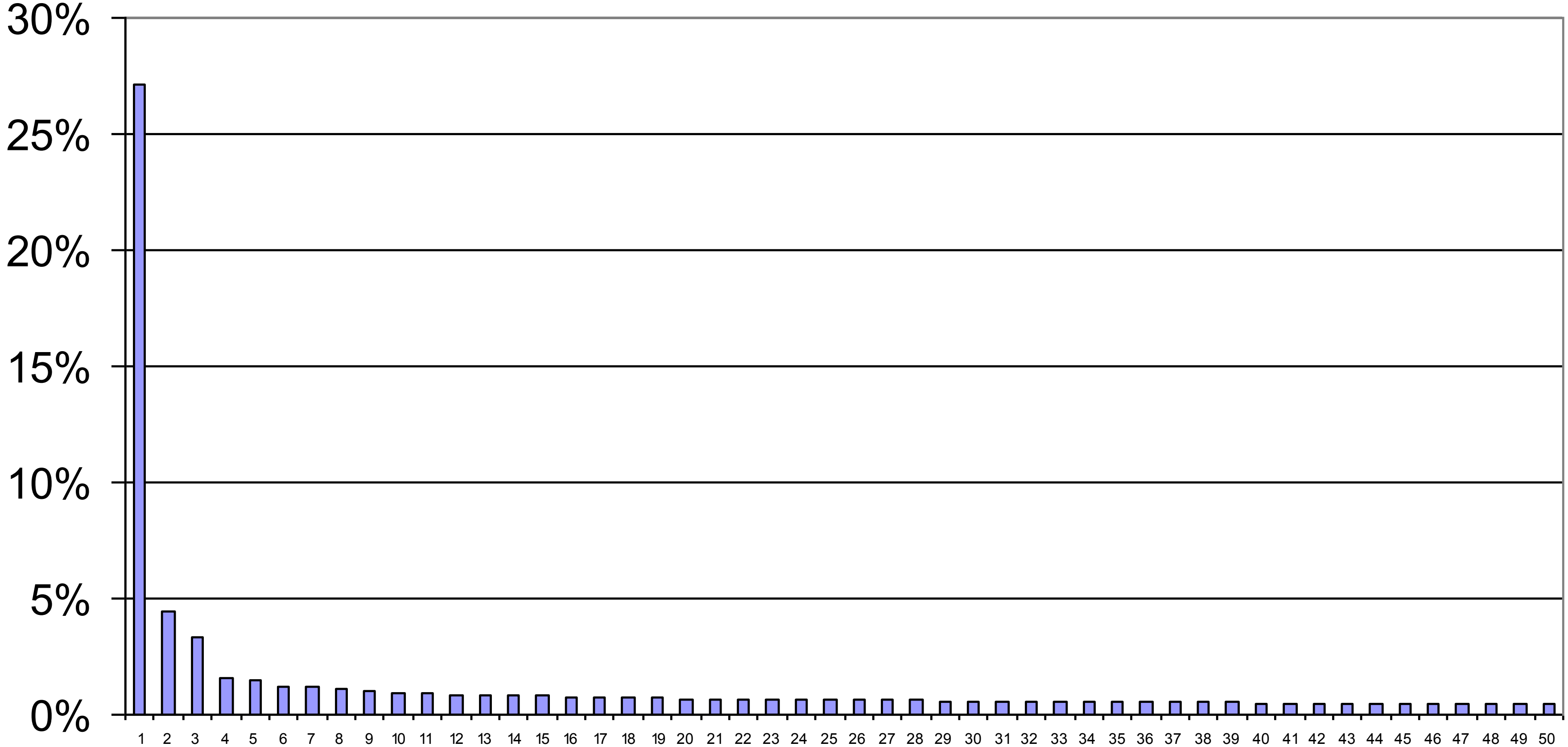
Then

$$R_i = \sum_{k=1}^{N} \beta_{\mathrm{ik}} F_k + \alpha \quad \text{expresses the result of multiple-regression of returns on factors}$$

50 largest eigenvalues using the 1400 US stocks with cap >1BB cap (Jan 2007)

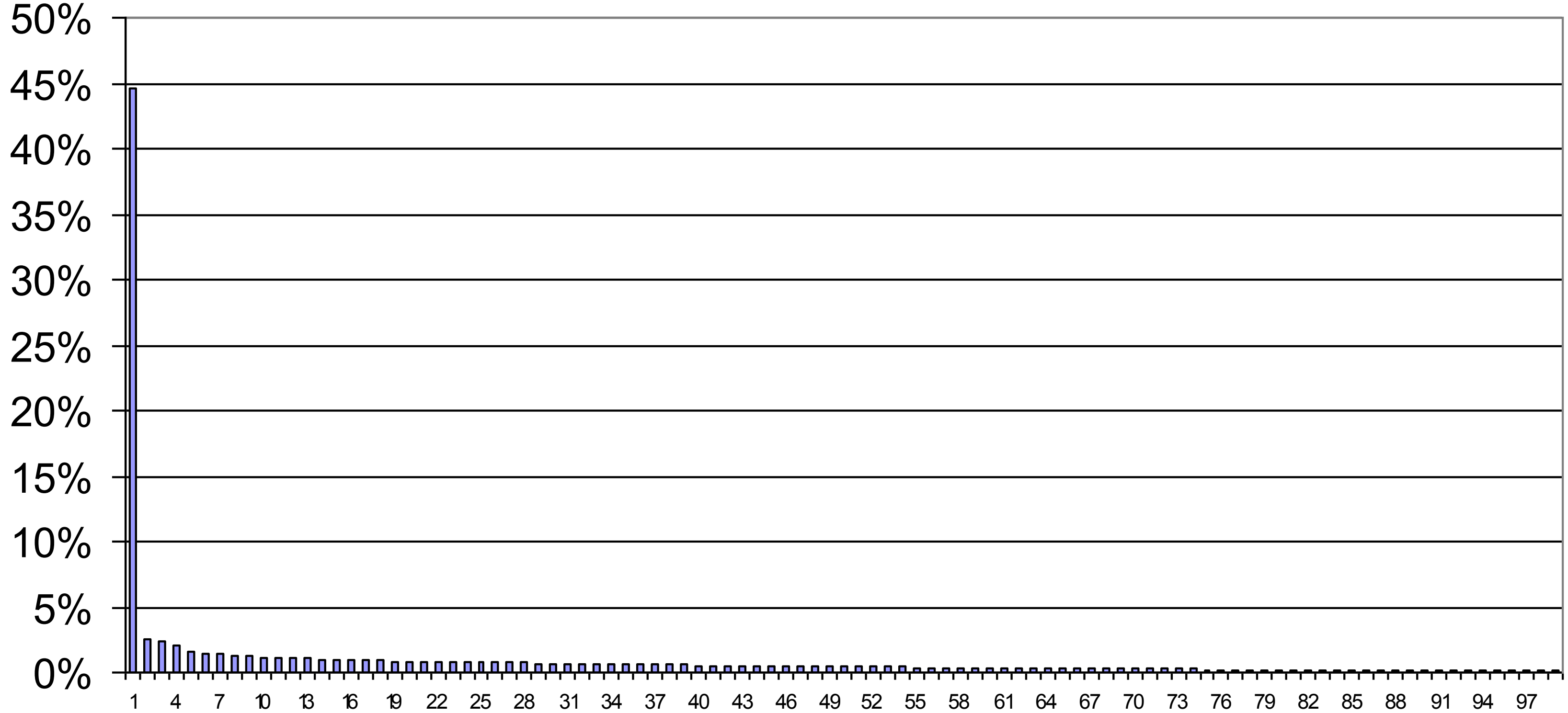


Top 50 eigenvalues for S&P 500 index components, May 1 2007,T=252



Nasdaq-100
Components of NDX/QQQQ

Data: Jan 30, 2007 to Jan 23, 2009
502 dates, 501 periods
99 Stocks (1 removed) MNST (Monster.com), now listed in NYSE



Bai and Ng 2002, *Econometrica*
Parsimonious approach for factor selection

$$I(m) = \min_{\alpha,\beta} \frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} \left(R_{it} - \sum_{k=1}^{m} \beta_{ik} F_{kt} - \alpha_i \right)^2$$

Least squares
penalty function

$$m^* = \arg\min_m \left(I(m) + m \cdot g(N,T) \right)$$

$$\lim_{N,T\to\infty} g(N,T) = 0, \quad \lim_{N,T\to\infty} \min(N,T) g(N,T) = \infty$$

Under reasonable assumptions on the underlying model, Bai and Ng prove that under PCA estimation, m^* converges in probability to the true number of factors as $N,T \to \infty$

Connection with eigenvalues of correlation matrix

$$J(m) \equiv \arg\min_{\alpha,\beta} \frac{1}{NT} \sum_{t=1}^{T} \sum_{i=1}^{N} \frac{1}{\sigma_i^2} \left(R_{it} - \sum_{k=1}^{m} \beta_{ik} F_{kt} - \alpha \right)^2$$

$$J(m) = \sum_{k=m+1}^{N} \lambda_k \qquad \text{also,} \quad I(m) = \sum_{k=m+1}^{N} \lambda_k \left(\sum_{i=1}^{N} \sigma_i^2 \left(V_i^{(k)} \right)^2 \right)$$

$$m^* = \arg\min_m \left(\sum_{k=m+1}^{N} \lambda_k + m g(N,T) \right)$$

Linear penalty function

For finite samples, we need to adjust the slope *g(N,T).*
Apparently, Bai and Ng (2002) tend to underestimate the number of factors in Nasdaq stocks considerably. (**2 factors**, T=60 monthly returns, N=8000 stocks)

Useful quantities

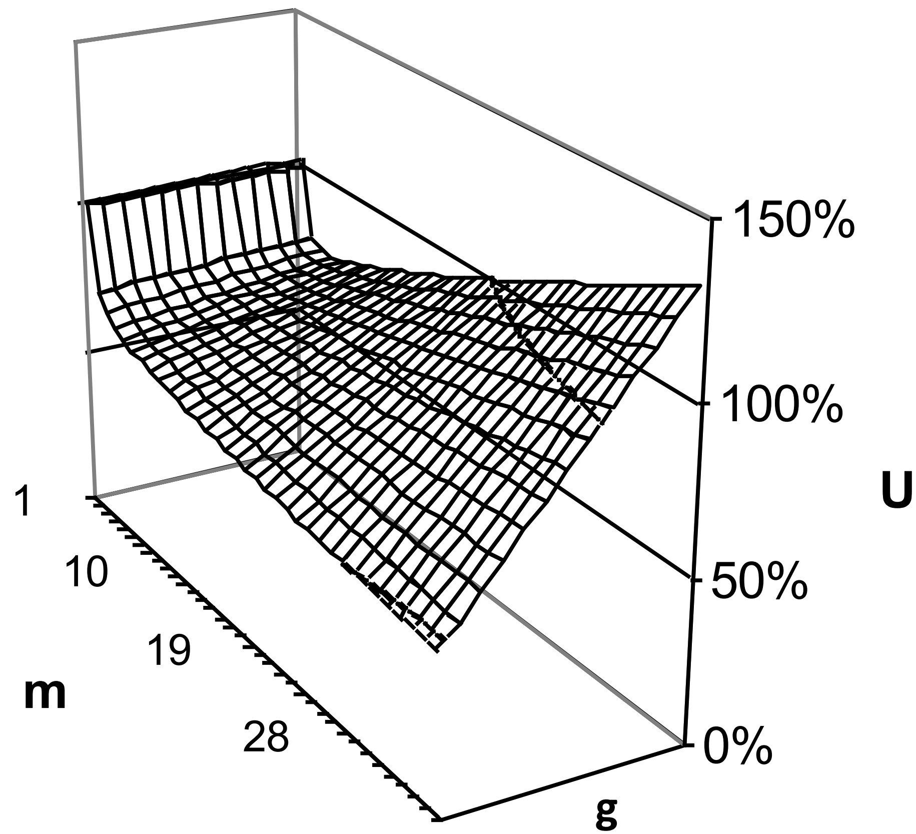
$$\frac{1}{N}\sum_{k=1}^{m}\lambda_k = \text{ Explained variance by first } m \text{ eigenvectors}$$

$$\frac{1}{N}\sum_{k=m+1}^{N}\lambda_k = \text{Tail}$$

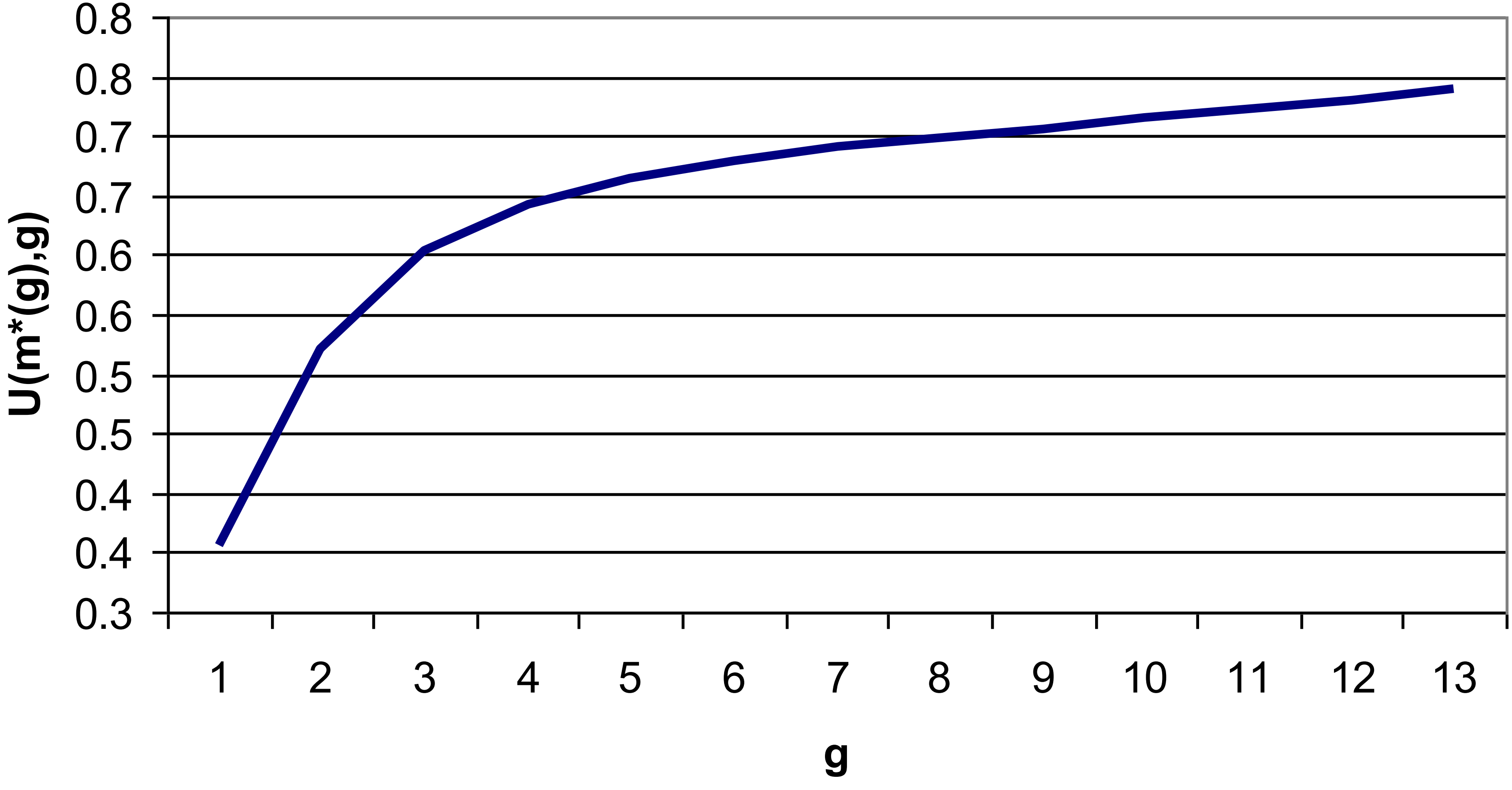
$$\frac{1}{N}\sum_{k=m+1}^{N}\lambda_k + g\,\frac{m}{N} = \text{Objective Function} = U(m,g)$$

$$\text{Convexity} = \frac{\partial^2 U(m^*(g),g)}{\partial g^2}$$

Objective function U(m,g)



Optimal value of U(m,g) for different g



Implementation of Bai & Ng on SP500 Data

g	m*	Lambda_m*	Explained Variance	Tail	Objective Fun	Convexity
1	117	0.20%	87.88%	12.12%	0.355	-
2	59	0.39%	71.44%	28.56%	0.522	-0.085085
3	29	0.59%	57.11%	42.89%	0.603	-0.041266
4	16	0.76%	48.51%	51.49%	0.643	-0.018110
5	10	0.96%	43.52%	56.48%	0.665	-0.007000
6	7	1.18%	40.43%	59.57%	0.680	-0.003096
7	6	1.22%	39.25%	60.75%	0.691	-0.004872
8	4	1.56%	36.56%	63.44%	0.698	0.001069
9	4	1.56%	36.56%	63.44%	0.706	0.000000
10	4	1.56%	36.56%	63.44%	0.714	0.000000
11	4	1.56%	36.56%	63.44%	0.722	0.000000
12	4	1.56%	36.56%	63.44%	0.730	0.000000
13	4	1.56%	36.56%	63.44%	0.738	-

If we choose the cutoff m* as the one for which the sensitivity to g is zero, then m*~5 to 7 seems appropriate.
This would lead to the conclusion that the S&P 500 corresponds to a 5-factor model.
The number is small in relation to industry sectors and to the amount of variance explained by industry factors.

The density of states: a useful formalism for finding significant EVs

Spectral theory as seen by physicists – origins in Quantum Mechanics and High Energy Physics.

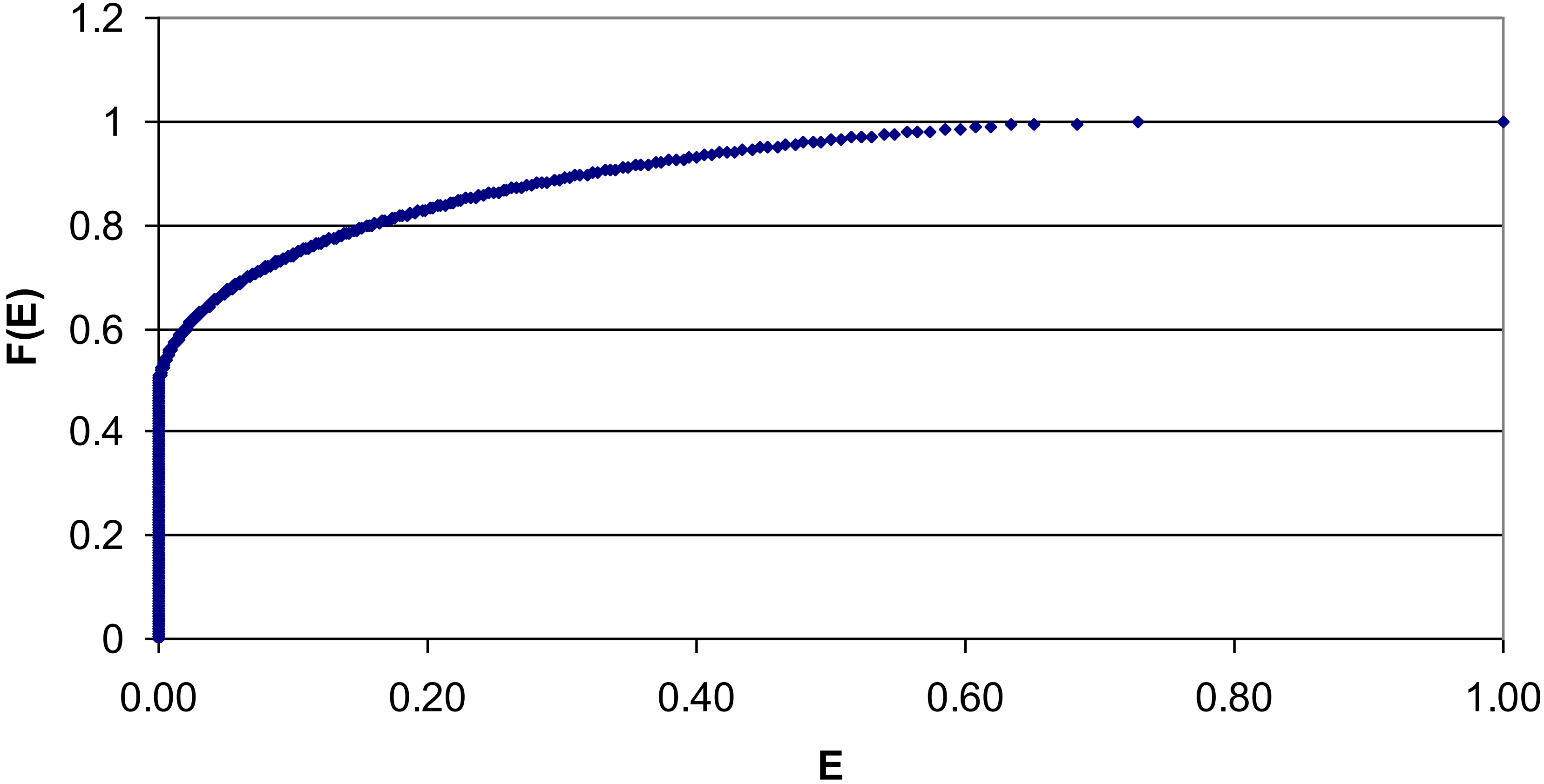
$$F(E) \equiv \frac{\#\{k : \lambda_k / N \leq E\}}{N} \qquad F(E) \text{ is increasing}, \quad F(1) = 1$$

$$f(E) = \frac{1}{N} \sum_k \delta\left(E - \frac{\lambda_k}{N}\right) \quad \therefore \quad F'(E) = f(E) \quad \text{D.O.S.}$$

One way to think about the DOS is as changing the x-axis for the y-axis,i.e. counting the number of eigenvalues in a neighborhood of any *E, 0<E<1.*

Intuition: if N is large, the eigenvalues of the insignificant portion of the spectrum will ``bunch up'' into a continuous distribution *f(E).*

Integrated DOS



In the DOS language…

$$\frac{1}{N}\sum_{k=m+1}^{N}\lambda_k = \int_0^{\lambda_m} E\, f(E)\,dE, \qquad \frac{m}{N} = 1 - F(\lambda_m)$$

$$U(E,g) = \int_0^{E} x\, f(x)\,dx + g\left(1 - F(E)\right)$$

$$\frac{\partial U(E,g)}{\partial E} = E\, f(E) - g f(E) = (E-g) f(E)$$

If $f(g) \neq 0$, then $E^*(g) = g$.

Dependence of the problem on g

$$V(g) = U\left(E^*(g), g\right) = \int_0^g xf(x)dx + g\left(1 - F(g)\right)$$

$$= gF(g) - \int_0^g F(x)dx + g - gF(g)$$

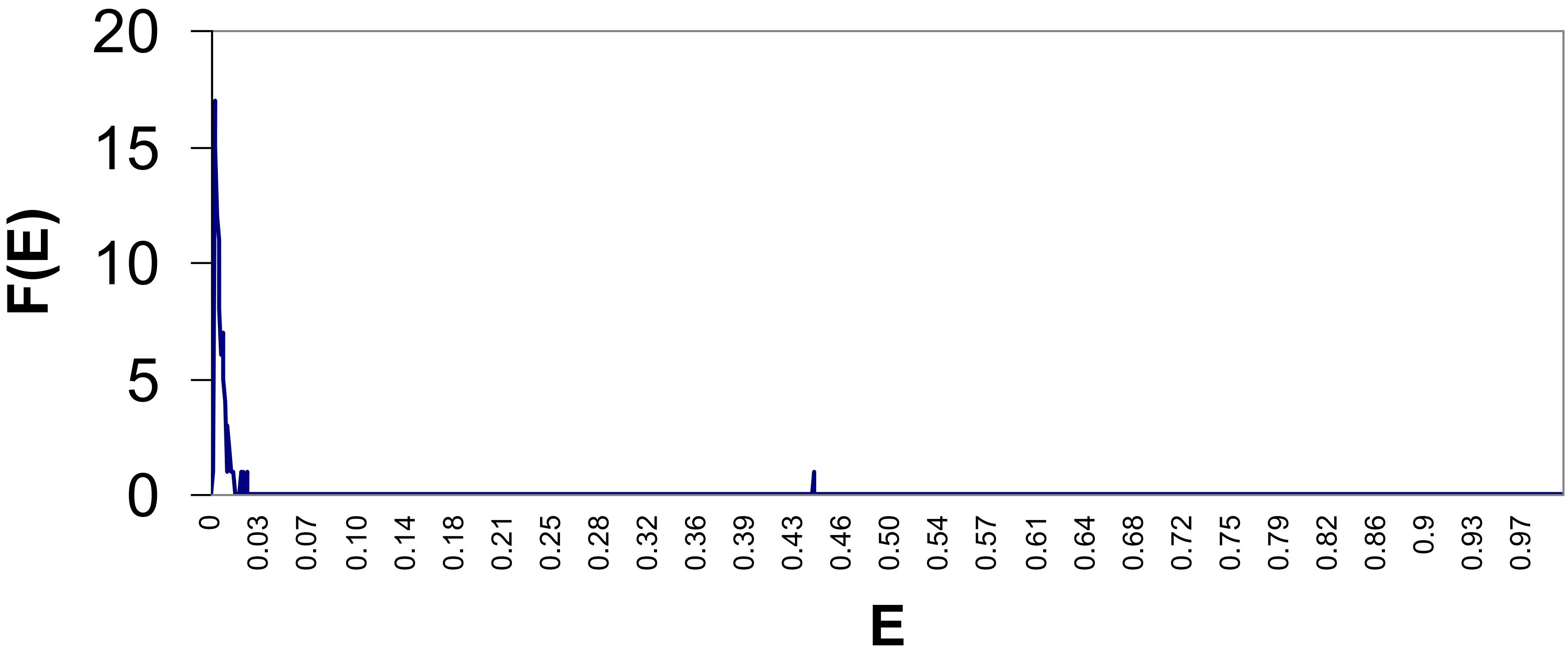
$$= g - \int_0^g F(x)dx$$

$$V'(g) = 1 - F(g)$$

$$V''(g) = -f(g)$$

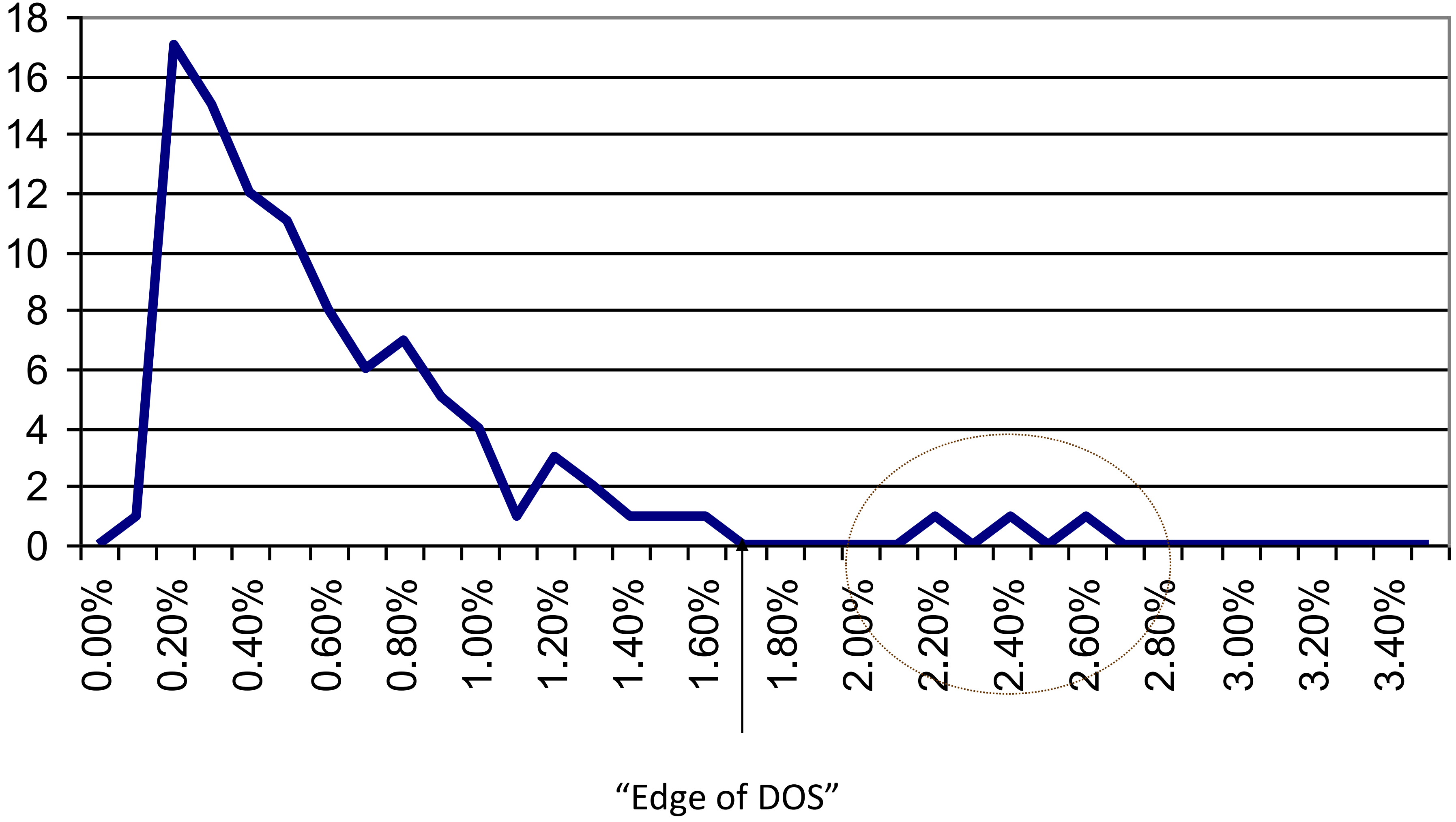
According to this calculation, the best cutoff is the level E where the DOE vanishes (or nearly vanishes) coming from the left, i.e. from the smallest eigenvalues.

Density of States (from previous data for Nasdaq 100)



One large mass at 0.44,
Some masses near 0.025
Nearly continuous density for lower levels

Zoom of the DOS for low eigenvalues



Random Matrix Theory

$$X_{tn},\ t=1,2,..,T,\quad n=1,2,...,N$$

$$X \sim N(0,1)$$

$$W_{mn}=\sum_{t=1}^{T} X_{tm}X_{tn},\qquad \mathbf{W}=\mathbf{X}^{\mathbf{t}}\mathbf{X}$$

$$\lambda_n,\ n=1,2,...N\qquad \text{eigenvalue s of } \mathbf{W}$$

What are the statistical properties of the eigenvalues as *N, T* tend to infinity?

What are the fluctuations of the eigenvalues for large *N, T*?

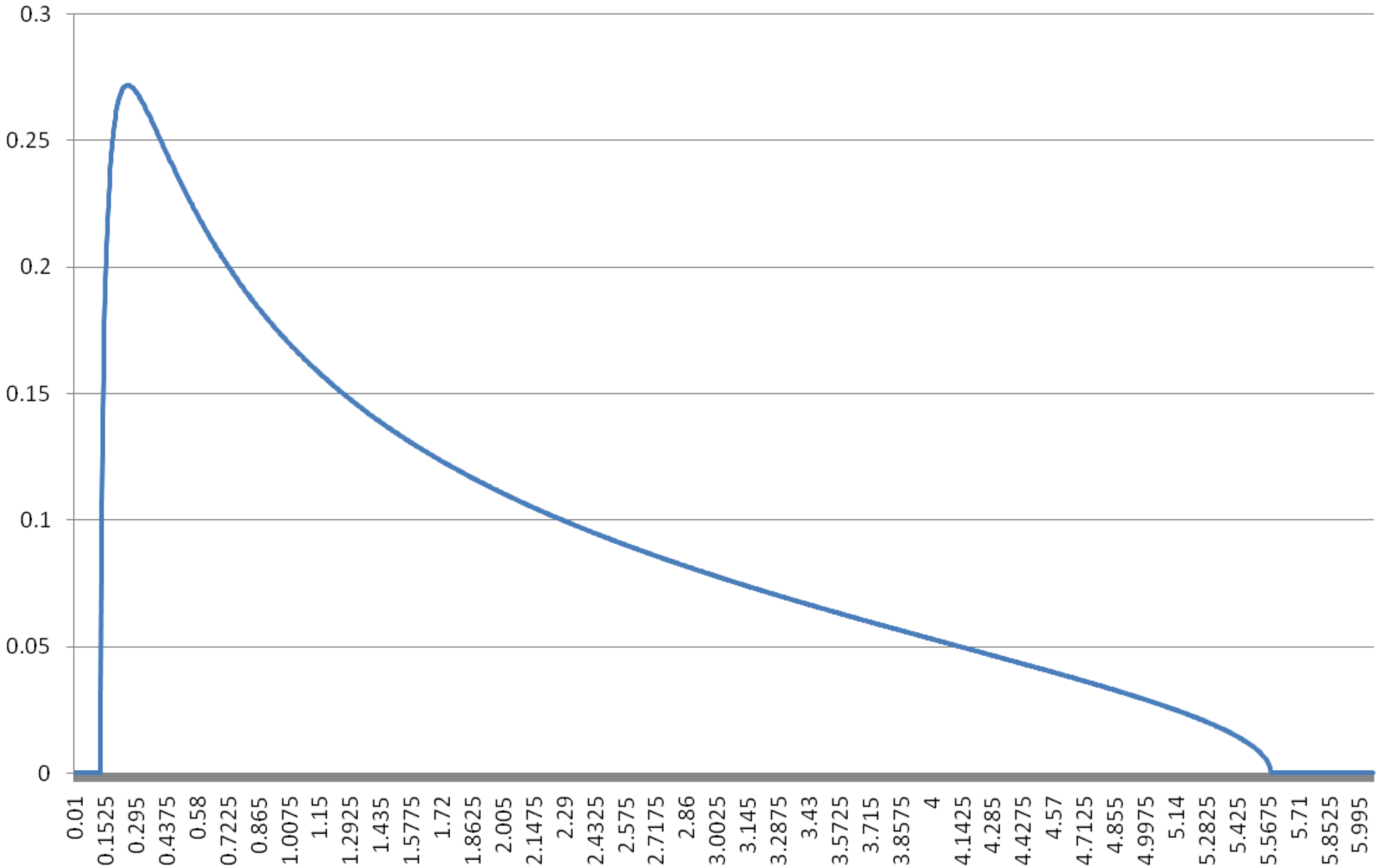
Marcenko-Pastur Distribution for the DOS of a Random Correlation Matrix

Theorem: Let *X* be a *T* by *N* matrix of standardized normal random variables and let *C=X'X*. Then, the DOS of C approaches the Marcenko Pastur distribution as *N,T* tend to infinity with the ratio *N/T* held constant.

$$\gamma = \frac{N}{T} \qquad \lambda_+ = \left(1+\sqrt{\gamma}\right)^2 \qquad \lambda_- = \left(1-\sqrt{\gamma}\right)^2$$

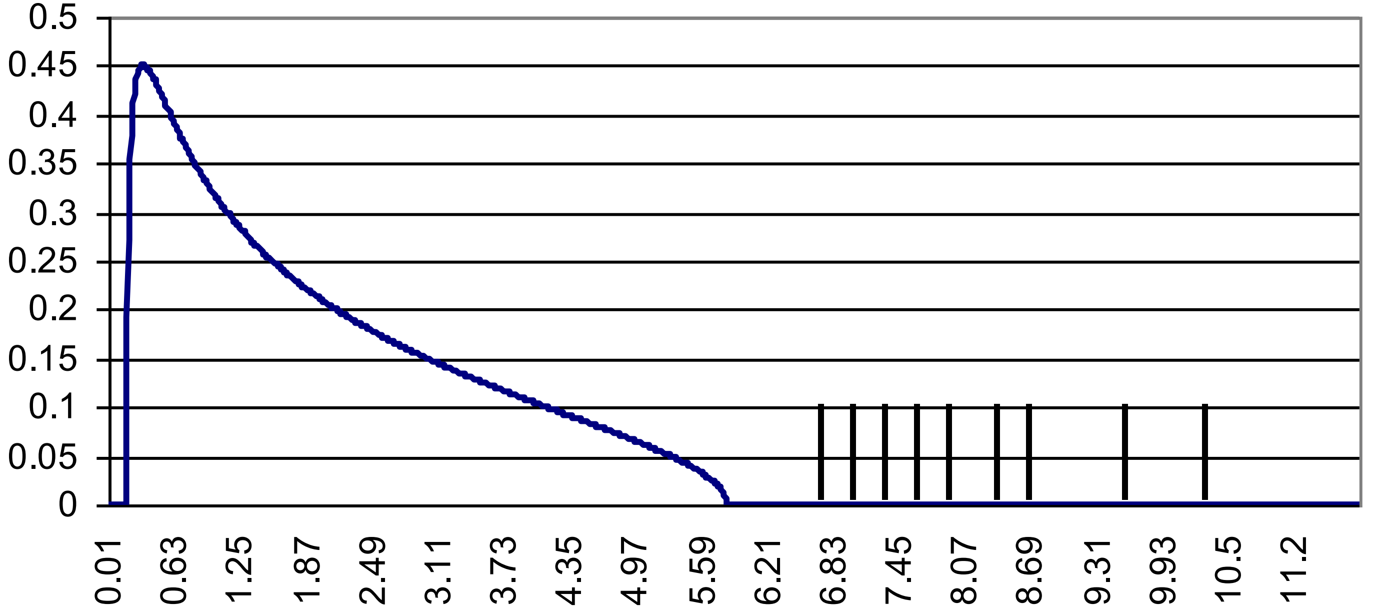
$$MP(\lambda) = \left(1-\frac{1}{\gamma}\right)^+ \delta(\lambda) + \frac{1}{2\pi\gamma}\frac{\sqrt{(\lambda_+-\lambda)(\lambda-\lambda_-)}}{\lambda}$$

Marcenko-Pastur Distribution
gamma=500/269= 1.858736



Bouchaud, Cizeau, Laloux, Potters (PRL, 1999)

The bulk distribution for spectrum of S&P 500
is described approximately by Marcenko-Pastur properly normalized but there
are detached eigenvalues (the significant ones!)



Bulk spectrum   significant eigenvalues

Results of PCA with DOS analysis for Nasdaq 100

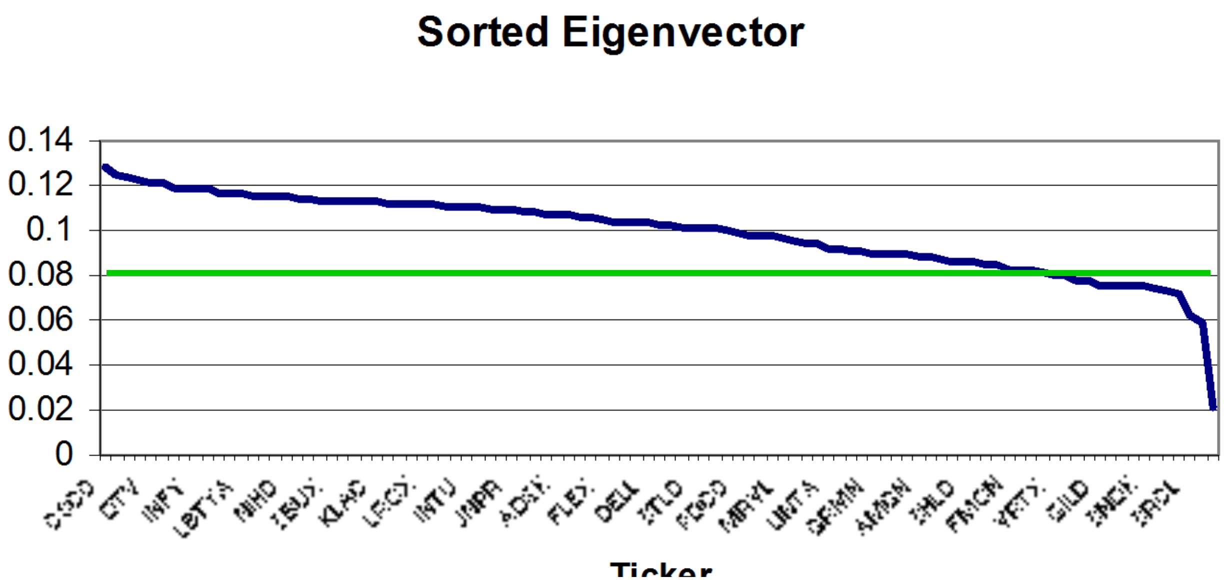
-- 4 significant eigenvectors/eigenvalues

-- first Eigen-state explains about 44% of the correlation

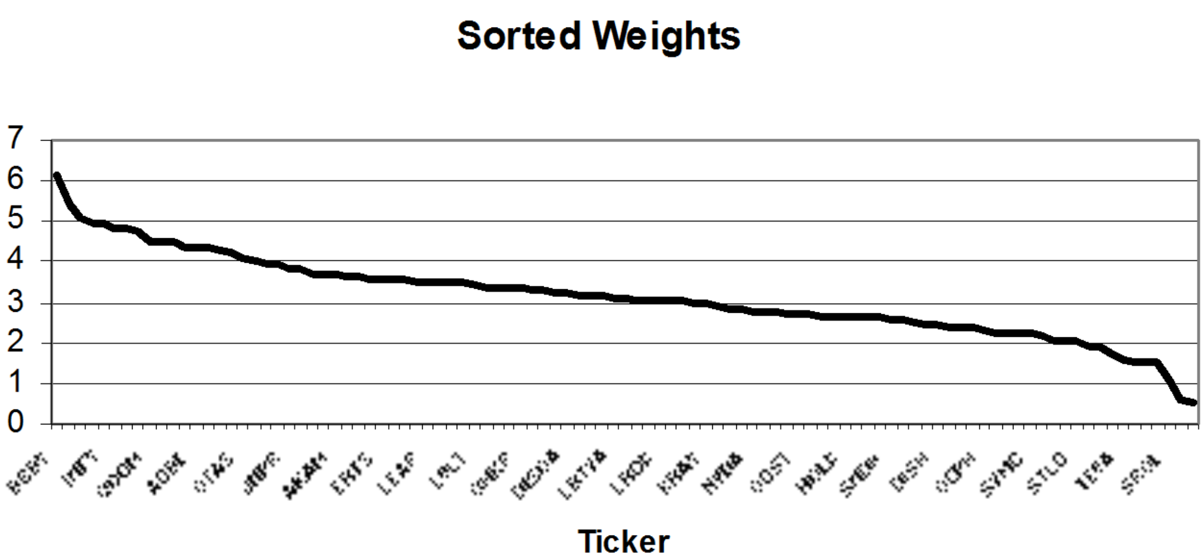
-- total explained variance= 51%

-- Now we need to indentify the eigenportfolios in terms of real market factors (industry, size, etc, etc).

First Eigenvector: Market

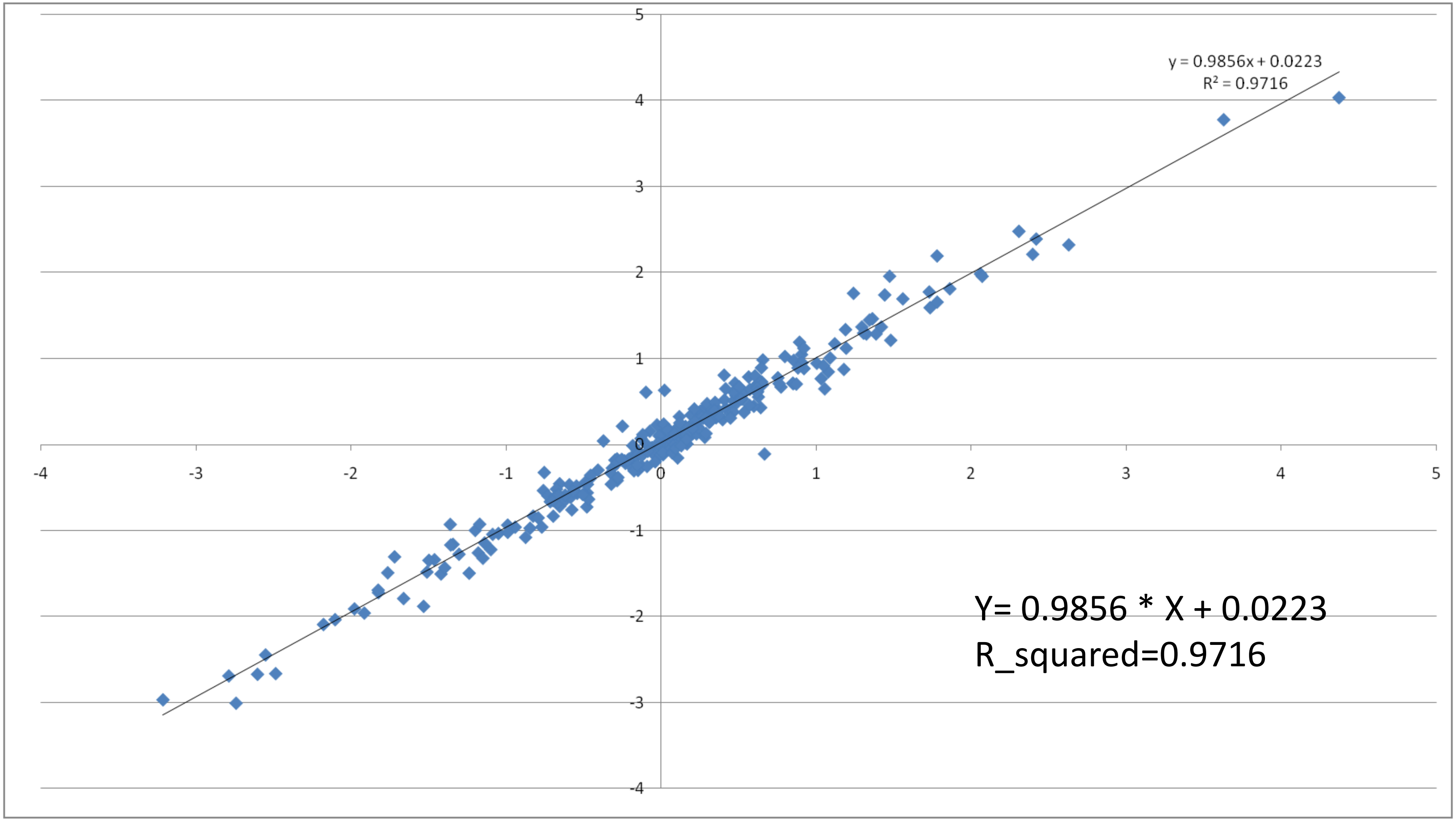


CSCO
INTC
ORCL
ADBE
DTV
SIAL
MSFT
SPLS
INFY
PAYX

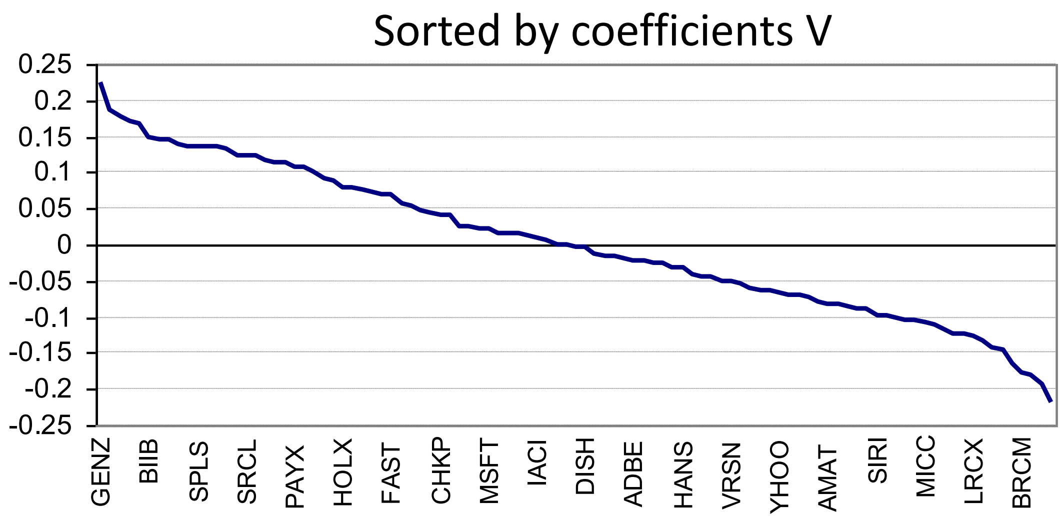


BBBY
LLTC
DELL
MRVL
INFY
ISRG
CMCSA
CTXS
QCOM
CSCO

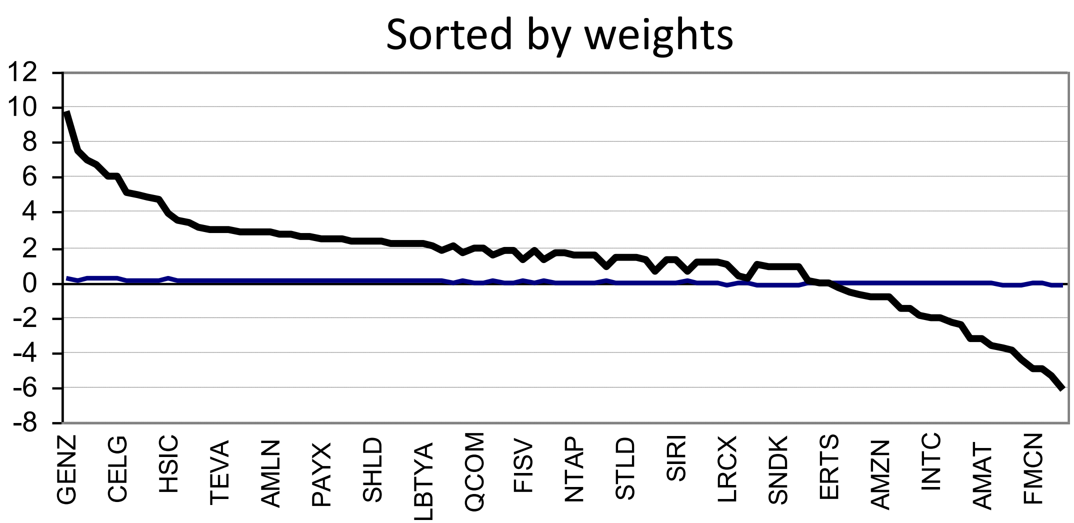
Returns of First PCA eigenportfolio (S&P 500) compared with S&P 500 returns (1/5/2009 to 1/29/2010)



Second Eigenvector: Biotech vs. Chips

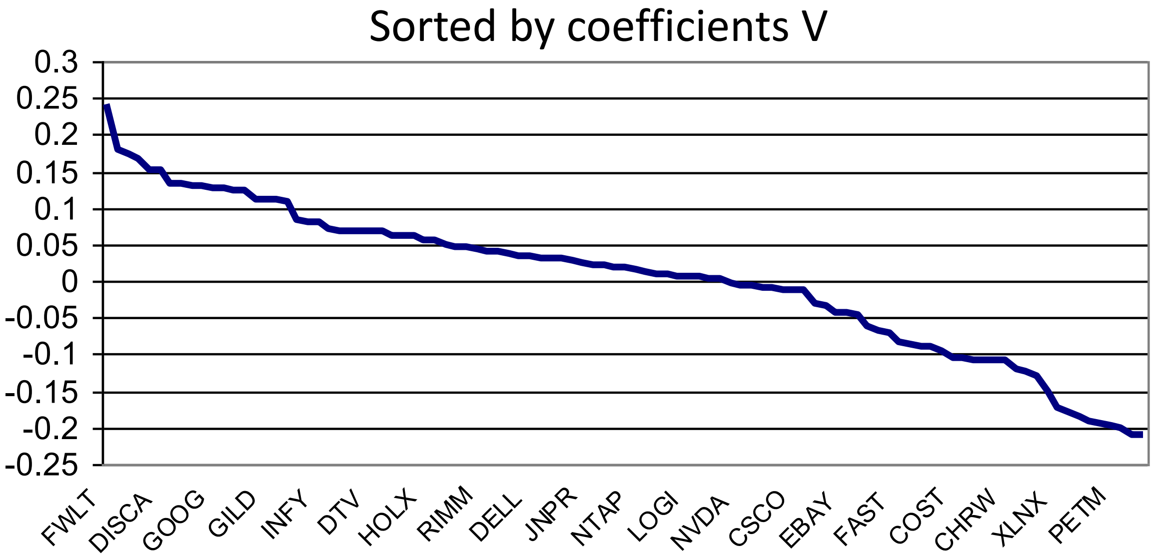


Top 10	Bottom 10
GENZ	MRVL
CEPH	NVDA
HSIC	FWLT
CELG	BRCM
GILD	SNDK
BIIB	JOYG
XRAY	RIMM
AMLN	BIDU
CTAS	LRCX
ESRX	ALTR



Top 10	Bottom 10
GENZ	BRCM
BBBY	FWLT
BIIB	DELL
GILD	FMCN
CEPH	AKAM
CELG	BIDU
ESRX	ALTR
CTAS	FLEX
AMGN	AMAT

Third Eigenvector: Manufacturing vs. Chips



Top 10	Bottom 10
FWLT	KLAC
LEAP	ALTR
JOYG	BBBY
STLD	PETM
TEVA	LRCX
DISCA	AMAT
DISH	LLTC
CEPH	SHLD
ATVI	XLNX
LBTYA	BRCM

Top 10	Bottom 10
LEAP	BBBY
FWLT	LLTC
DISCA	SHLD
FMCN	AMAT
NIHD	ALTR
ATVI	PETM
GILD	MRVL
CEPH	LRCX
GOOG	BRCM
JOYG	KLAC

``Coherence''

Definition: If an eigenvector is such that stocks with a given property (size, industry sector) have entries with the same sign, then the eigenvector is said to be coherent (with respect to the given property).

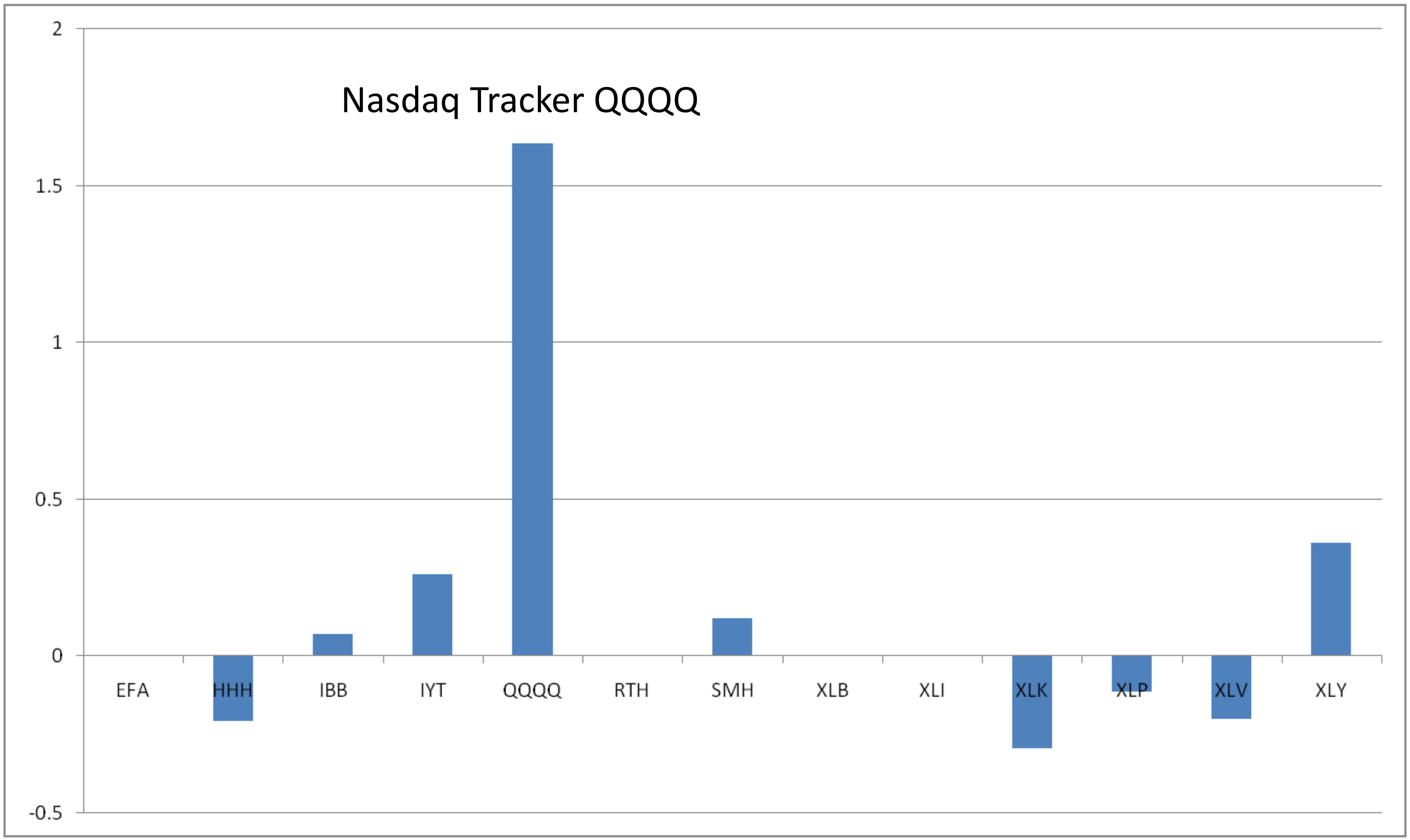
Conjecture: The significant eigenvectors are coherent with respect to either size of sector

Identification of the Eigenportfolios via ETFs

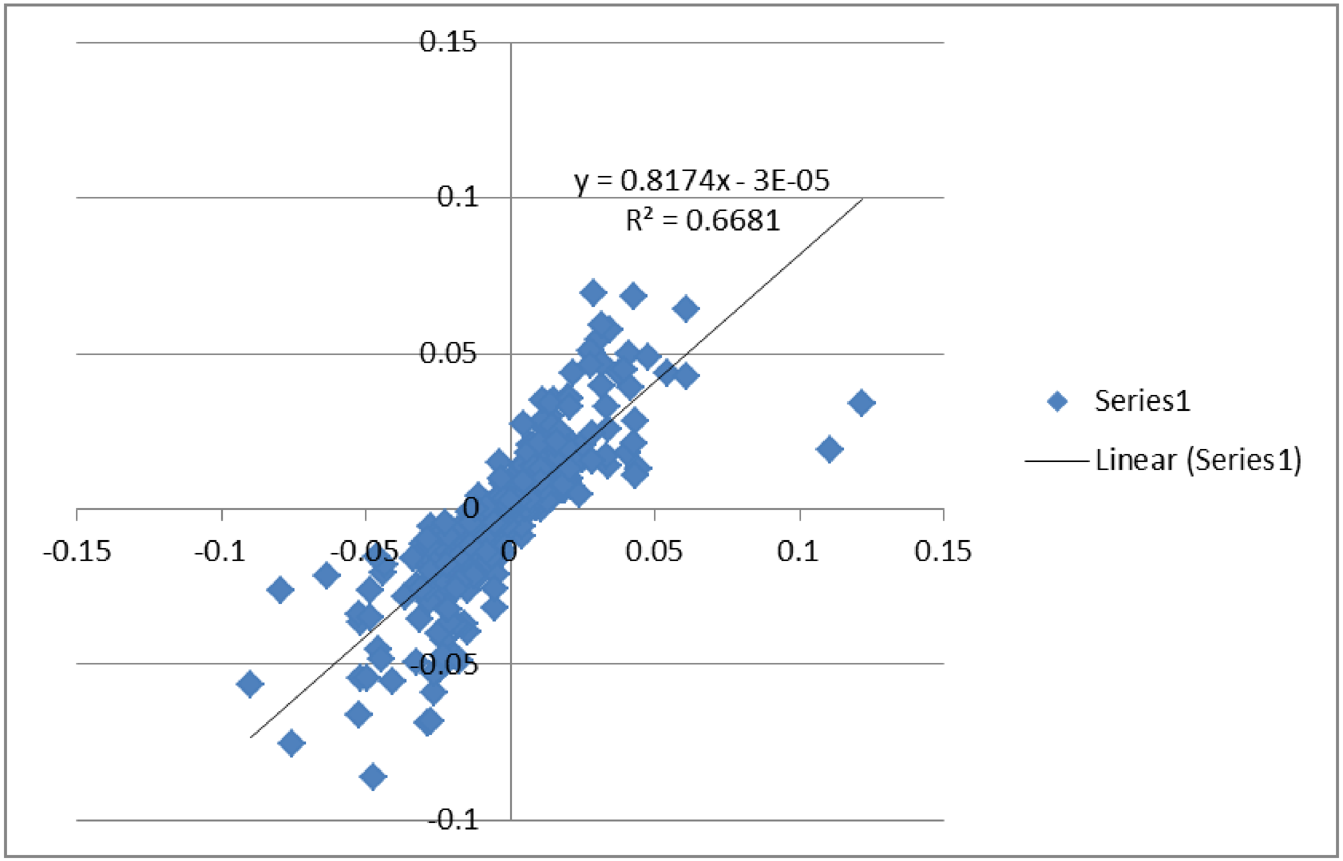
Indentify the Eigenportfolios by making multiple regressions or ``greedy'' regressions on the returns of Exchange Traded Funds

EFA	Europe & Far East
HHH	Internet
IBB	Biotechnology
IYT	Transportation
QQQQ	Nasdaq 100 Index Tracker
RTH	Retail
SMH	Semiconductors
XLB	Materials
XLI	Industrials
XLK	Technology
XLP	Consumer Staples
XLV	Health Care
XLY	Consumer Discretionary

First Eigenportfolio (NDX)



Scaled returns QQQQ vs 1st EV



Second Eigenportfolio (NDX)